

#guardsymposium2026
X @GuardConsortium



GUARD **INTERNATIONAL** **SYMPOSIUM**

GU-Alliance for Research and Development

PRESENCIAL RETRANSMITIDO EN DIRECTO
FACE-TO-FACE AND LIVE STREAMING



16-18 SEPTIEMBRE 2026

Espacio Maldonado, Madrid

Explainable Machine Learning for Modeling Metastatic Risk in Prostate Cancer

MET-score

**Sonia Maciá Escalante, MD, PhD
SMED Clinical Research**

01 · CONCEPTUAL FRAMEWORK

What are Big Data and Artificial Intelligence?

And why they are transforming precision oncology

From data to knowledge: Big Data, Machine Learning and Explainability

What is Big Data?

- Thousands of samples, millions of genetic variables
- Impossible to analyze manually
- 766 tumors × 16,234 genes = 12.4 million data points

What does Machine Learning do?

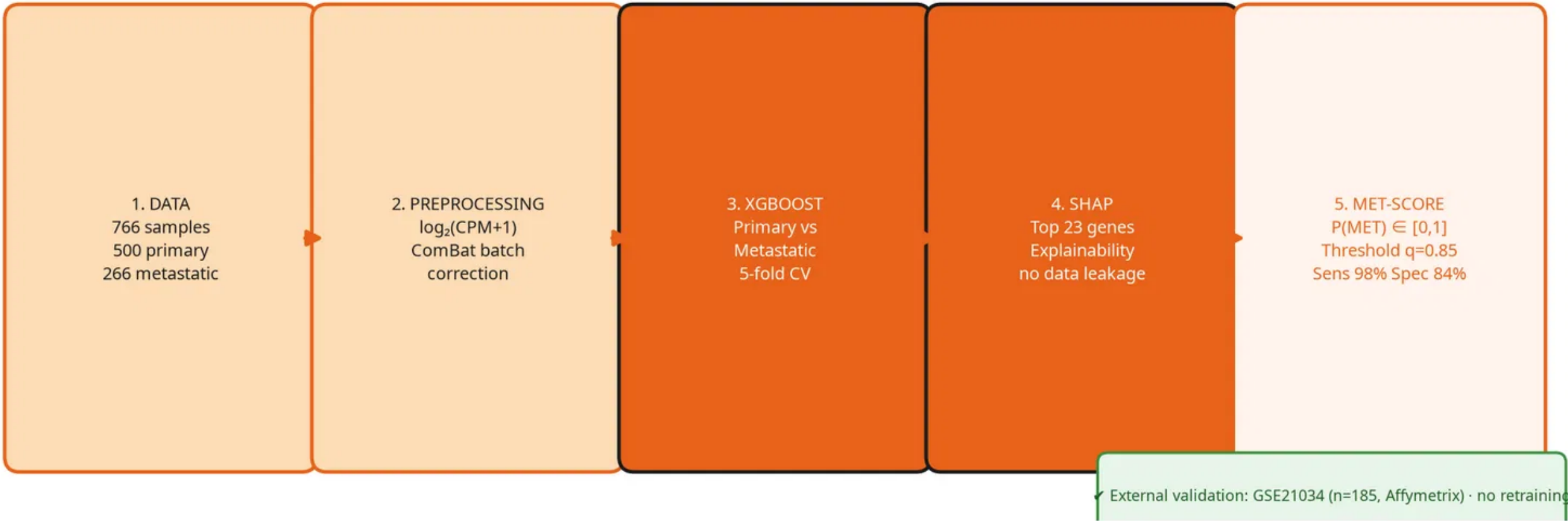
- Learns hidden patterns without explicit instructions
- Classifies, predicts and quantifies risk
- Explainable ML (SHAP) opens the "black box": which genes it uses and how much weight it gives them

02 · THE CLINICAL PROBLEM

What molecularly differentiates a primary tumor from a metastasis?

Previously published signatures had failed to resolve this

Methodological design: from 16,234 genes to 23 key genes



Primary and metastatic prostate tumors differ profoundly in their transcriptome. However, previously published signatures had critical limitations that prevented their robust clinical use.

Data: TCGA-PRAD + SU2C · 500 primary · 266 metastatic · 16,234 genes

Model: XGBoost with L1+L2 regularization · stratified 5-fold CV · 2,000 bootstraps

Explainability: SHAP re-selected per fold (no data leakage)

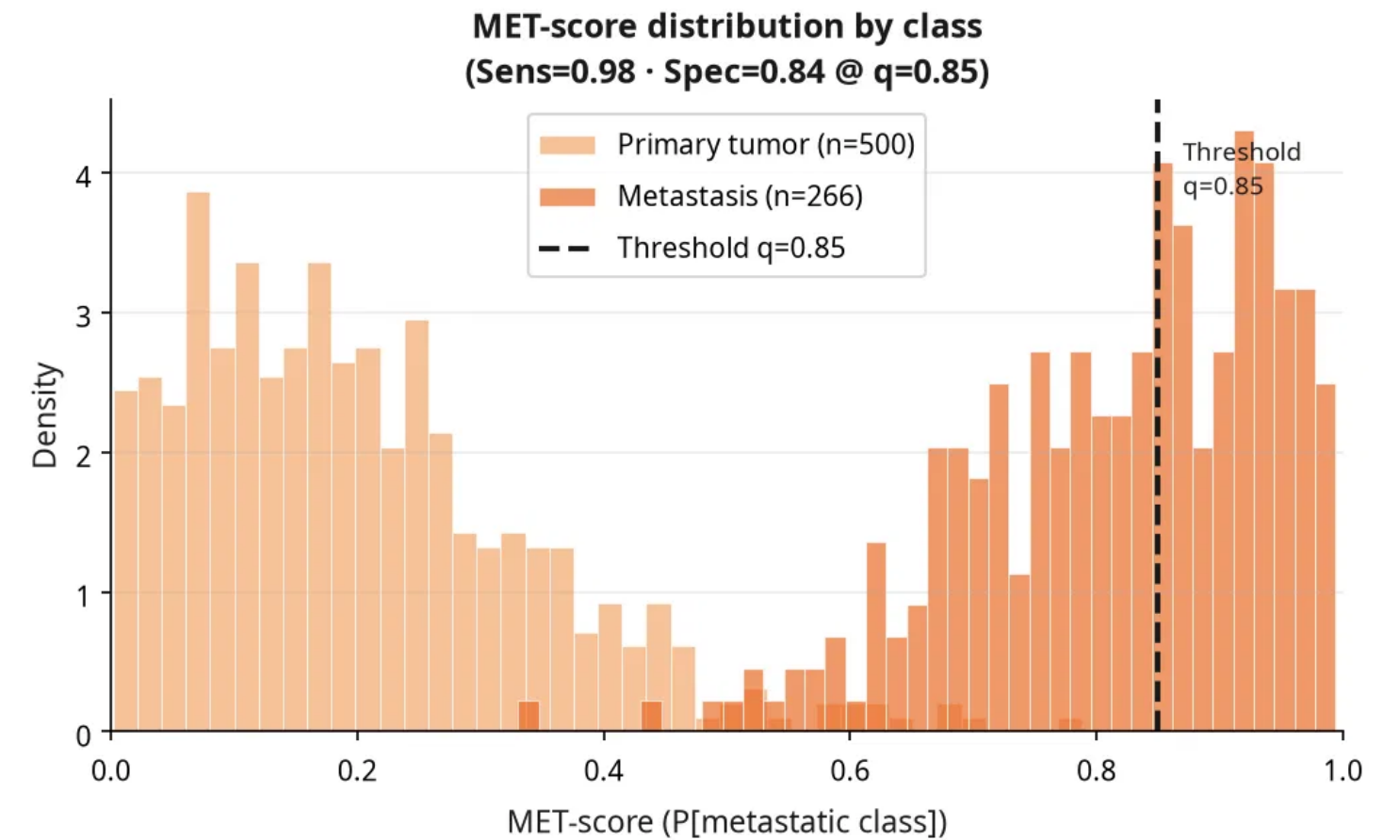
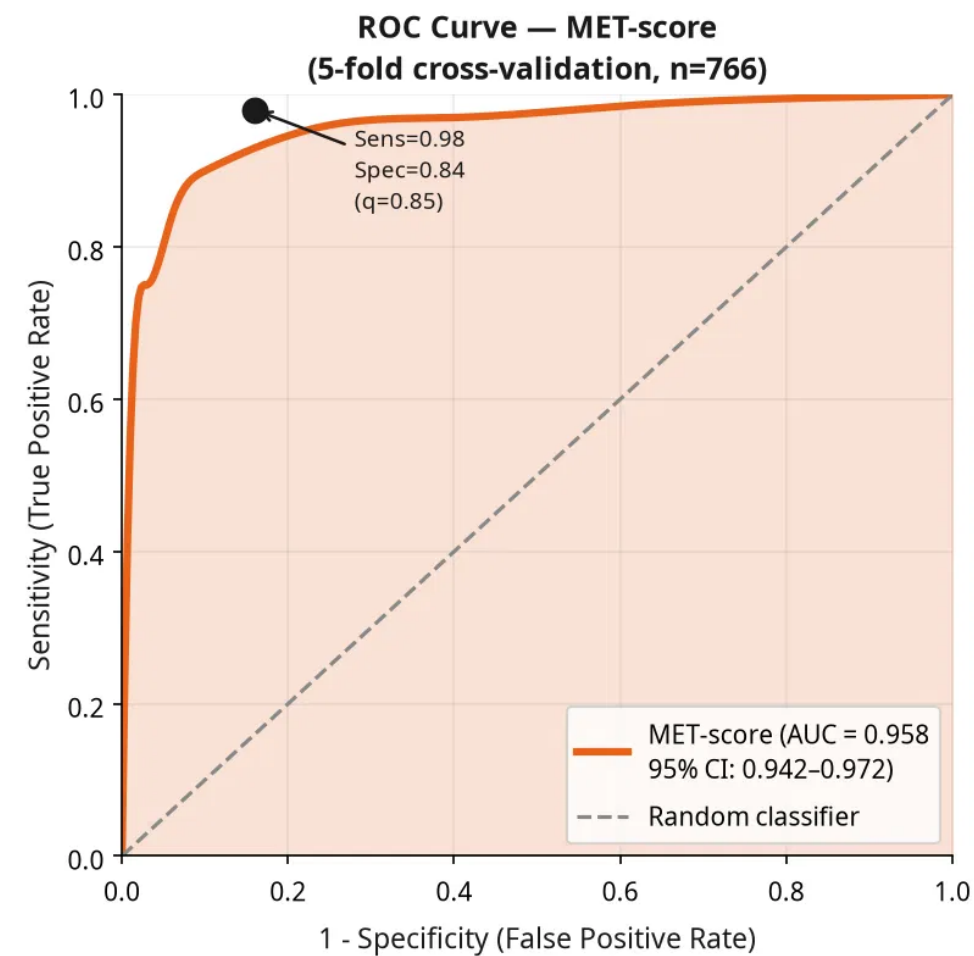
External validation: GSE21034 (Affymetrix, n=185) · no retraining

03 · RESULTS

Exceptional performance, real biological coherence

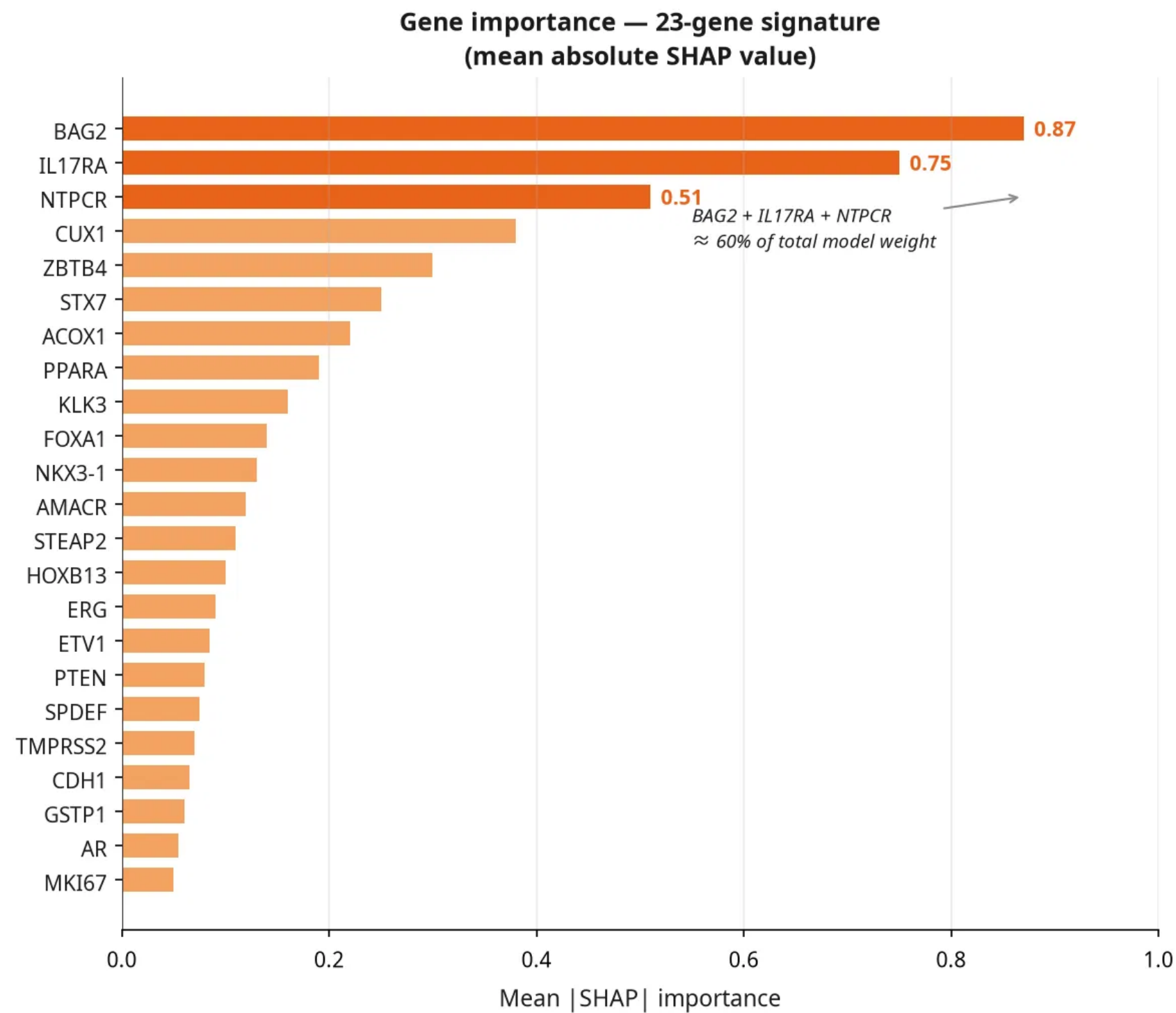
From an AUC of 0.958 to a validated 23-gene signature

AUC 0.958: the MET-score discriminates with high accuracy



An AUC of 1.0 is perfection. An AUC of 0.5 is a coin toss. **0.958 is very close to perfection** — Sensitivity 98% · Specificity 84% at threshold q=0.85.

BAG2 · IL17RA · NTPCR dominate: 3 genes account for 60% of model weight



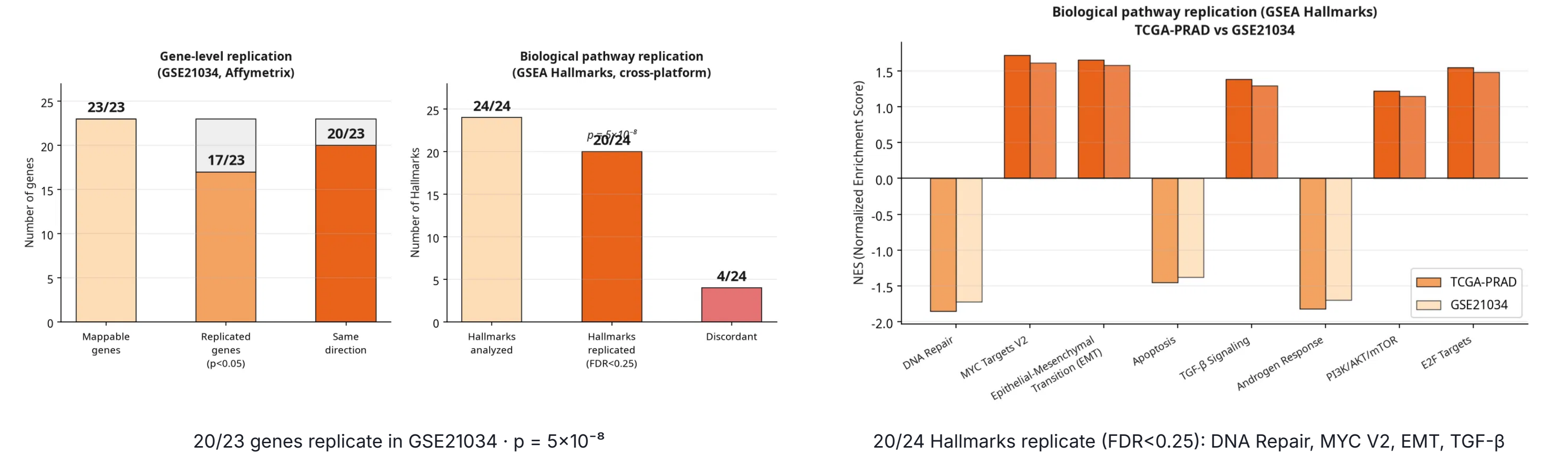
Of 16,234 genes analyzed, the model selected a signature of only **23 genes**.

The top 3 genes account for **60%** of the model's decision weight.

All signature genes are predominantly **underexpressed** in metastases.

Consistent with the **loss of prostatic lineage markers** during metastatic progression.

The signature replicates on an independent platform: real biological signal, not an artifact

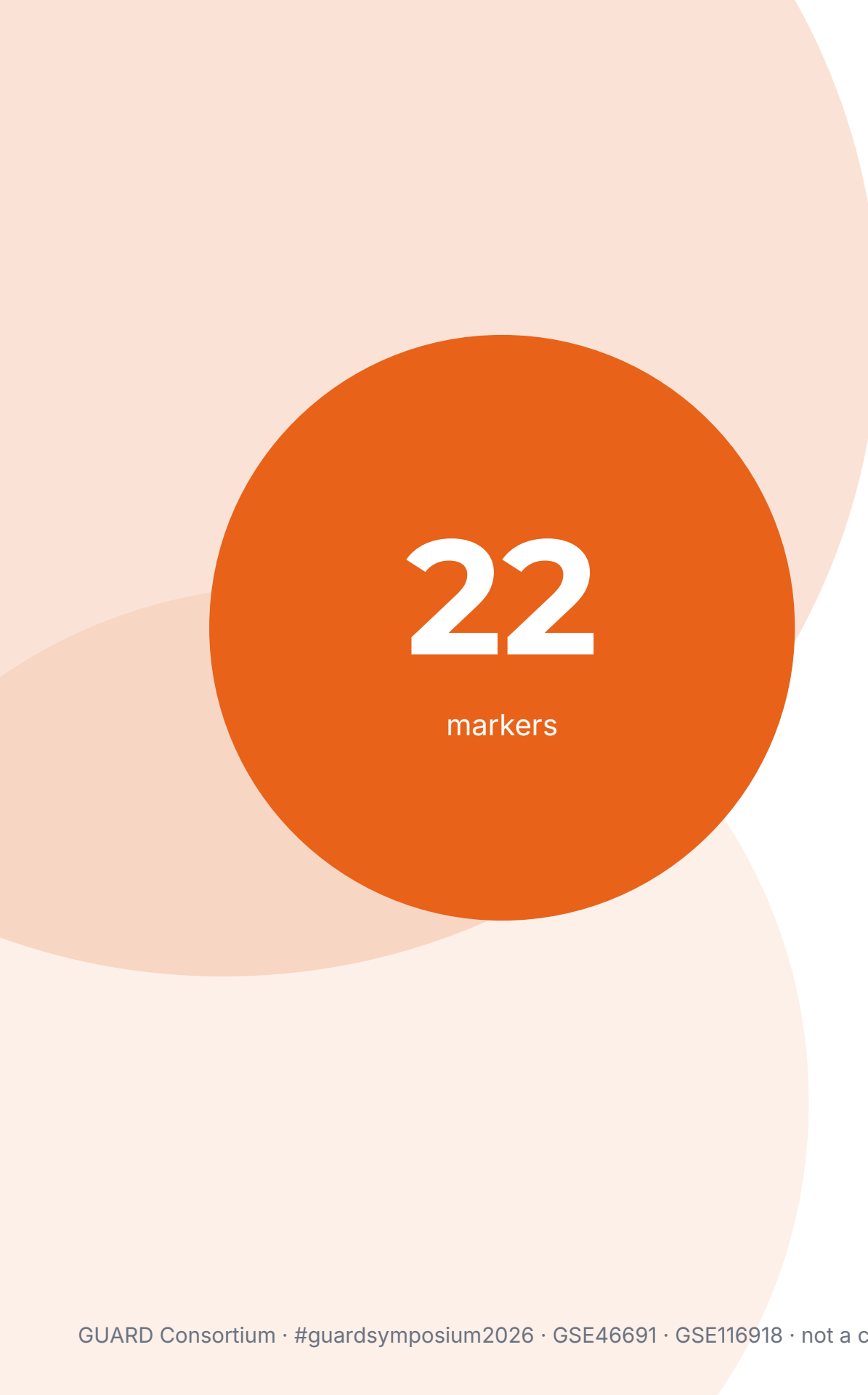


No retraining. Different platform. Same result. That is biological robustness.

Five key study conclusions

- 1 A 23-gene signature with **AUC 0.958** that distinguishes primary from metastatic tumors with high accuracy
- 2 Cross-platform replication: **20/23 genes** and **20/24 Hallmarks** replicate without retraining
- 3 Consistent biological signal: loss of prostatic lineage markers in metastases
- 4 Explainable Machine Learning: we know which genes the model uses and how much weight it gives them
- 5 Next clinical step: can the MET-score predict which primary patient will metastasize?

The MET-score is not just a statistical model —
it is a biological map of the metastatic process



22

markers

LOCALIZED PROSTATE CANCER · GENOMIC BIOMARKER

Predicting metastasis from the primary tumour

A genomic score that outperforms the Gleason grade and separates risk within each histological grade

Development: 545 prostatectomies (Mayo Clinic) · 212 metastases

External validation: 248 biopsies (Belfast) · 22 metastases, no retraining

01 · THE PROBLEM

**Gleason groups patients;
it does not discriminate within the group**

Can gene expression in the primary tumour anticipate metastasis better than the microscope?

Five in a hundred will metastasize; the microscope cannot say which

Gleason 7 accounts for half of all patients and mixes indolent with aggressive disease. The decision —monitor, operate, intensify— is made on the biopsy, where Gleason is even less accurate than on the surgical specimen.

Limitation of the current scenario	What this work contributes
Gleason groups, it does not rank within the group	Continuous score that reorders patients across categories
Decisions are made on biopsy, where grade is noisier	Validated precisely on diagnostic biopsies
Previous signatures with no cell-composition control	Stroma, immune infiltrate and endothelium regressed out
Marker selection performed on the full dataset	Selection redone inside every fold (out-of-fold)
Signatures of hundreds of genes, hard to translate	22 markers, measurable by qPCR or a dedicated array

Clinical consequence: patients who do not need treatment are overtreated, and those who do are undertreated. The goal is to estimate, before choosing treatment, how likely that particular patient is to develop metastasis.

Two independent cohorts, one single endpoint

The model is trained on the first cohort, frozen, and applied to the second with no refitting. Sample type, treatment and measurement technology all change at once.

Development — GSE46691 (Mayo Clinic)	Validation — GSE116918 (Belfast)
Radical prostatectomy specimen	Diagnostic biopsy
Radical surgery	Radical radiotherapy
545 patients · 212 metastases	248 patients · 22 metastases
Binary endpoint	Endpoint and time to event
Affymetrix Human Exon 1.0 ST	Almac Prostate DSA
1,411,399 probes across the expressed genome	Frozen model, zero refitting

The endpoint is confirmed metastasis — not biochemical recurrence or any other surrogate. That is the difference between measuring what matters and measuring what is convenient.

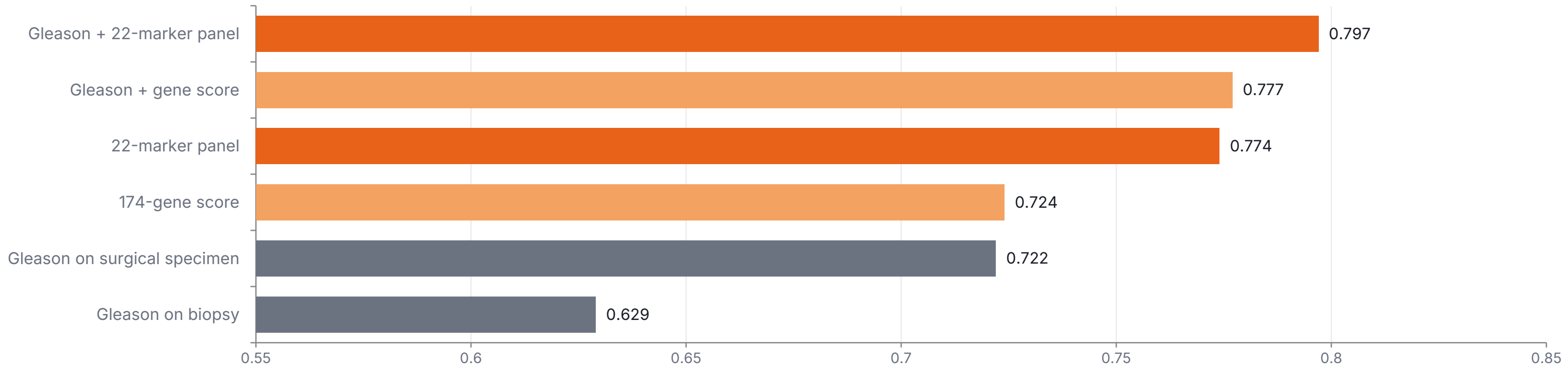
02 · RESULTS

The transcriptome beats Gleason and adds information of its own

AUC 0.797 combining Gleason and the panel, versus 0.722 for Gleason alone

Everything tested, compared on the same endpoint

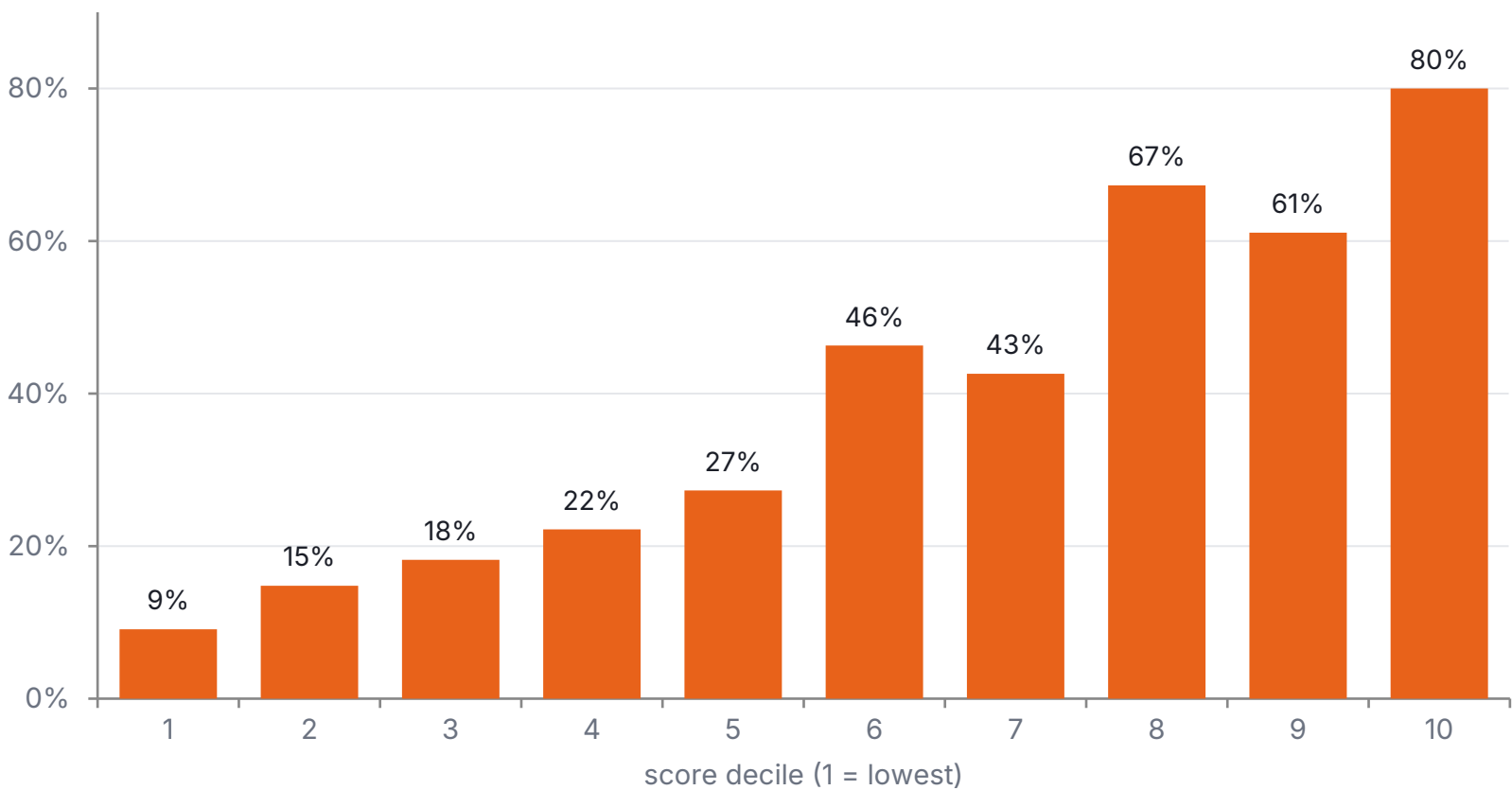
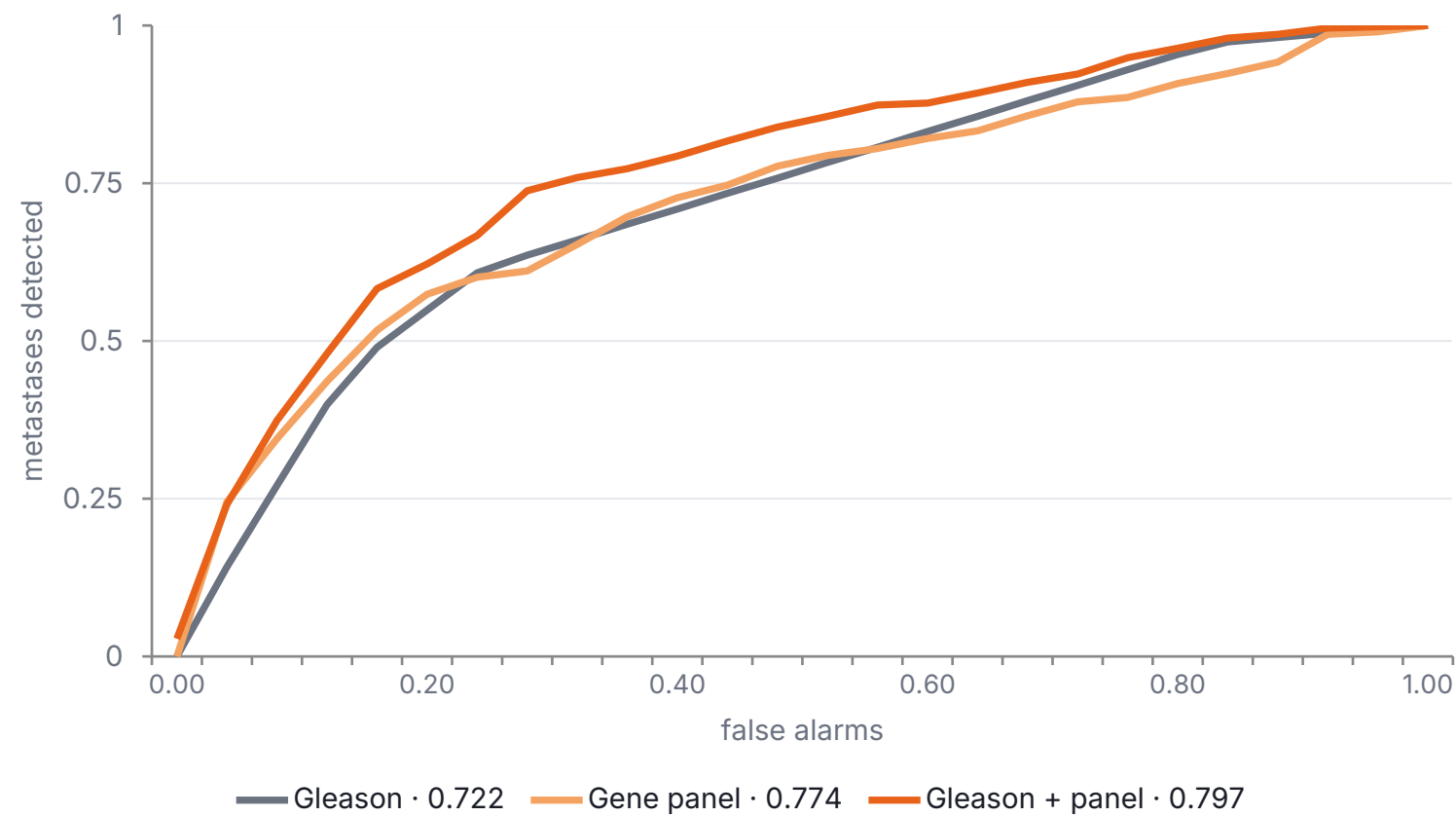
Ability to anticipate observed metastasis (AUC). A value of 0.50 is a coin toss; 1.00 would be perfect. No real biomarker gets there.



The improvement over Gleason is +0.075, with a confidence interval of [+0.050, +0.099]: it does not include zero. On a model that already contains Gleason, adding the score gives $p = 2.5 \times 10^{-14}$ — it is not reproducing grade by another route, it adds its own information.

Discrimination and observed risk in the development cohort

545 patients scored out-of-fold, 212 with documented metastasis.



With no formula in between, the observed rate of metastasis rises from 9% to 80% as the score increases. It is the most direct check that the score means something.

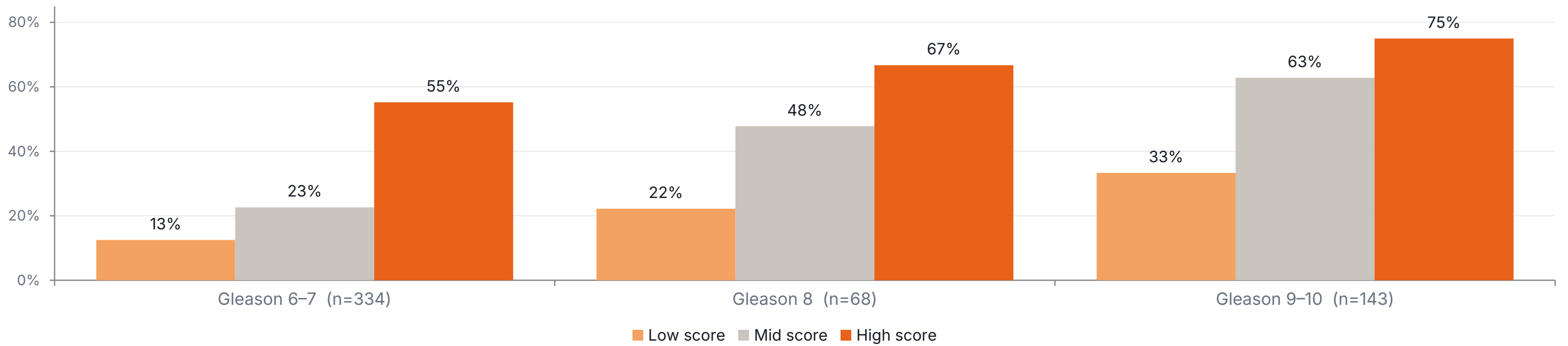
03 · CLINICAL UTILITY

It separates risk within each Gleason grade

From 12% to 55% metastasis within the same Gleason 6–7, with no visible difference under the microscope

Observed metastasis by Gleason grade and score

Actual percentage of patients who developed metastasis, splitting each Gleason group into thirds by genomic score. These are observed data, not predictions. If the score added nothing, the three bars within each block would be identical.



A Gleason 6-7 with a high score (55%) is worse off than the average Gleason 8 (48%); a Gleason 9-10 with a low score (38%) is better off. The score reorders patients across clinical categories, which is exactly what a marker must do to change decisions.

Thirty more metastases, without a single extra false alarm

AUC ranks patients, but it does not say how many are detected or how many are alarmed. Fixing the same cost for both options —alarming exactly 103 healthy people— lets us count in patients.

GLEASON ALONE

139 / 212

metastases detected

GLEASON + PANEL

169 / 212

at the same false-alarm cost

GAIN

+30

patients identified in time

FALSE ALARMS

103 / 103

identical under both options

Where it would fit in practice. The score is computed on the diagnostic biopsy, the only material available when treatment must be chosen. Its contribution is largest exactly where the decision is hardest: Gleason 7, which accounts for half of all patients and mixes indolent with aggressive disease.

Where it decides nothing. The value lies not in the AUC but in the separation of extremes: the top fifth and the bottom fifth of scores carry risks that differ several-fold. In the middle range the model decides nothing, and presenting it otherwise would be dishonest.

Using Gleason together with the score instead of Gleason alone is what turns 139 detected metastases into 169 out of the 212 in the cohort, without alarming anyone else.

04 · ROBUSTNESS

**It works outside where
it was built, and it is not chance**

A simultaneous triple change of sample, treatment and platform — with the model frozen

External validation: 248 biopsies, another hospital, another technology

The model was frozen and applied without refitting to patients treated with radiotherapy. It is the most demanding test in this work.

SCORE AUC

0.719

95% CI [0.597, 0.835]

GLEASON AUC

0.629

measured on biopsy

RISK MULTIPLIER

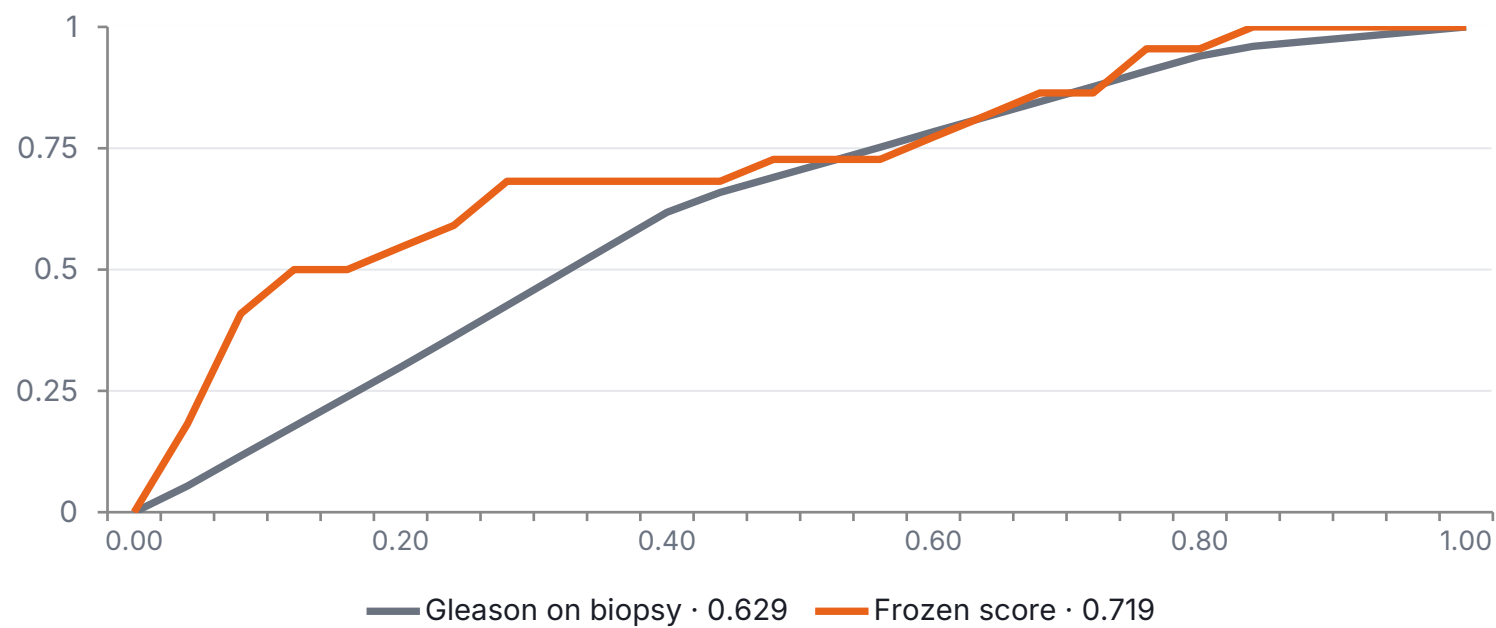
×2.34

per standard deviation of score

GLEASON ONCE ADJUSTED

p = 0.46

no longer significant



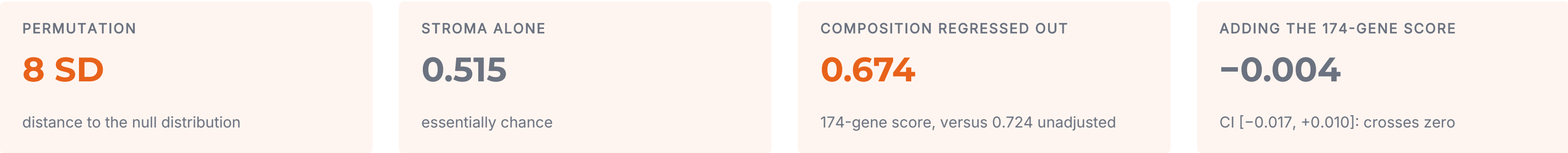
What changed all at once. Sample type (surgical specimen → biopsy), treatment (surgery → radiotherapy) and measurement technology. A model working only through quirks of its source cohort would not survive this triple change. Log-rank on time-to-metastasis curves: p = 0.0012.

A caveat that must be declared. Gleason scores 0.629 here versus 0.722 on the surgical specimen: part of the score's advantage reflects a comparator measured worse, because biopsy understages. That said, at the moment of decision the biopsy is all there is — and that is precisely the gap.

With 22 events the interval is wide, but it excludes chance.

Not chance, not tumour content, and it cannot be pushed higher

Two alternative explanations could produce these figures with no biology behind them. The permutation test shuffles the outcomes and repeats the whole pipeline, to see what the method yields when there is nothing to learn; the composition control regresses out smooth muscle, fibroblast, immune infiltrate and endothelium. Both are explicitly ruled out; then an improvement is actively searched for, and does not appear.

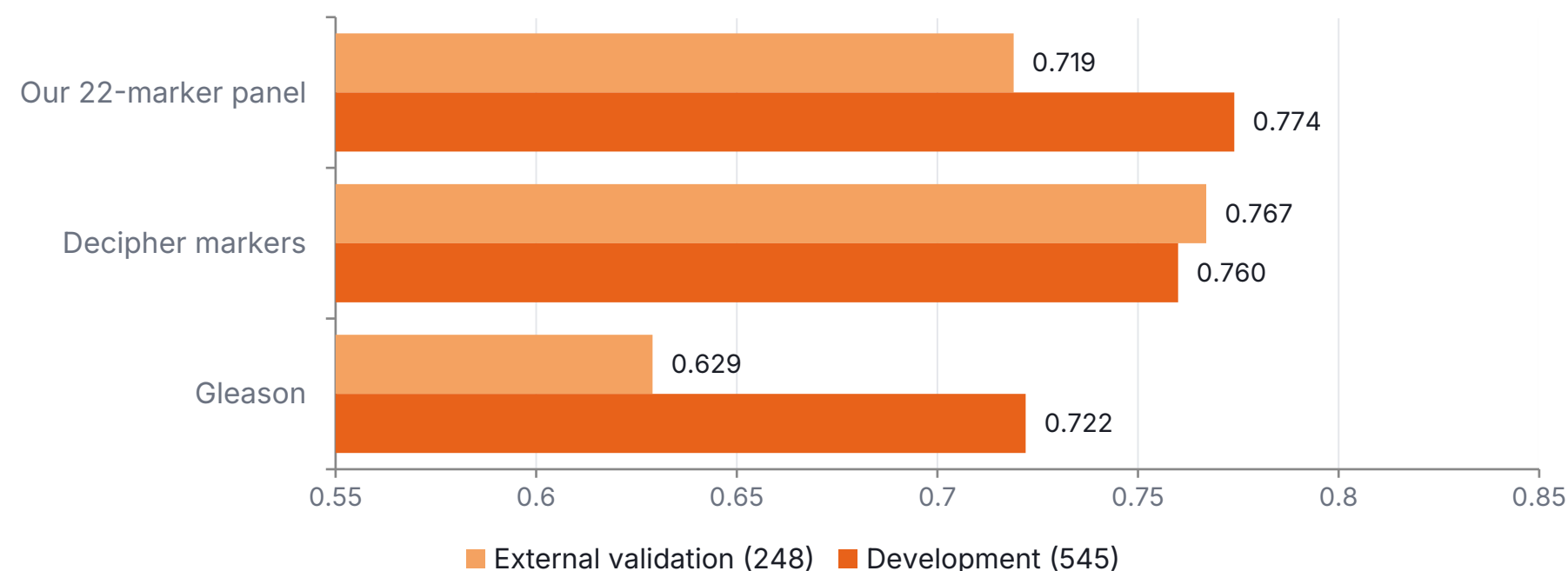


Combination tested	AUC · difference versus the reference
Gleason + panel (reference)	0.797 —
Gleason + panel + 174-gene score	0.793 [-0.017, +0.010]
+ composition-adjusted model	0.779 [-0.042, +0.005]
+ proliferation	0.775 [-0.041, -0.004]
Expanding the panel to 100 / 300 markers	flat to 100 markers, drops at 300

The ceiling is in the data, not in the modelling. Adding PSA, T stage and age does not help either: the score alone already gave 0.719, and with all four clinical variables it lands at 0.716. Breaking that ceiling would require another molecular layer —methylation, copy number— or more events.

Head to head with the reference classifier

Decipher is a genomic classifier already included in clinical guidelines, and its discovery cohort is the same one used here. The comparison is unavoidable, and the result is not the one you would hope for.



Independent convergence. The 22 markers map to fifteen annotated gene symbols, and seven of them appear among Decipher's published markers: ANO7, CAMK2N1, IQGAP3, MYBPC1, NUSAP1, TNFRSF19 and UBE2C. A selection starting from 1.4 million probes and never looking at their list recovered nearly a third of the commercial panel.

In development, a tie. +0.013 with an interval of $[-0.015, +0.044]$. Yet on a model already containing Gleason and Decipher, adding our panel still contributes ($p = 0.0006$): they are not the same score.

In external validation, slightly behind. 0.767 versus 0.719; the -0.048 difference has an interval that crosses zero. With 22 events they cannot be told apart.

The conclusion, unvarnished: our panel does not beat Decipher. It ties in development, sits slightly behind in external validation, and neither difference reaches significance. What it does show is that it carries partly independent information and that the selection procedure converges on the same markers from scratch.

Five conclusions

- 1 A 22-marker panel measured in the primary tumour predicts metastasis **better than the Gleason grade** (0.774 versus 0.722).
- 2 It carries information that is **independent of Gleason** and of the available clinical variables: $p = 2.5 \times 10^{-14}$ on a model that already contains it.
- 3 Translated into patients: **169 of 212 metastases detected versus 139**, without a single extra false alarm.
- 4 **It generalizes without retraining** to another cohort, another sample type, another treatment and another platform — and it is neither chance nor cell composition.
- 5 Its value lies where the decision is hardest: **Gleason 7 measured on biopsy**, which accounts for half of all patients.

This is not the leap from 0.72 to 0.99 promised by work that validates badly.

It is what holds up when you do not cheat.

Retrospective cohorts · 22 events in external validation · not a clinical tool.

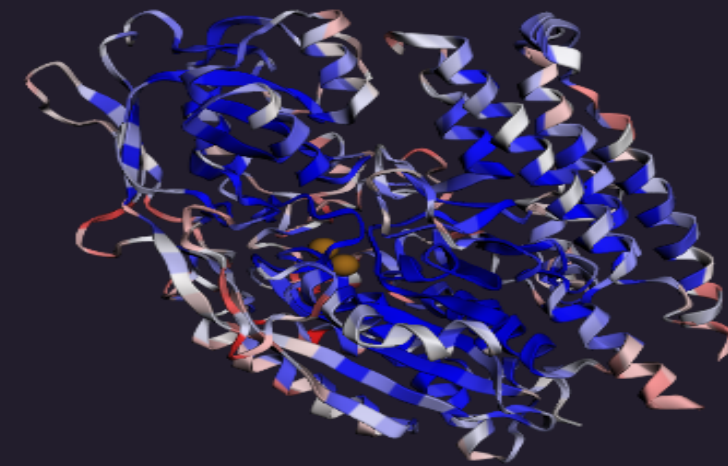


We design antibodies against PSMA.

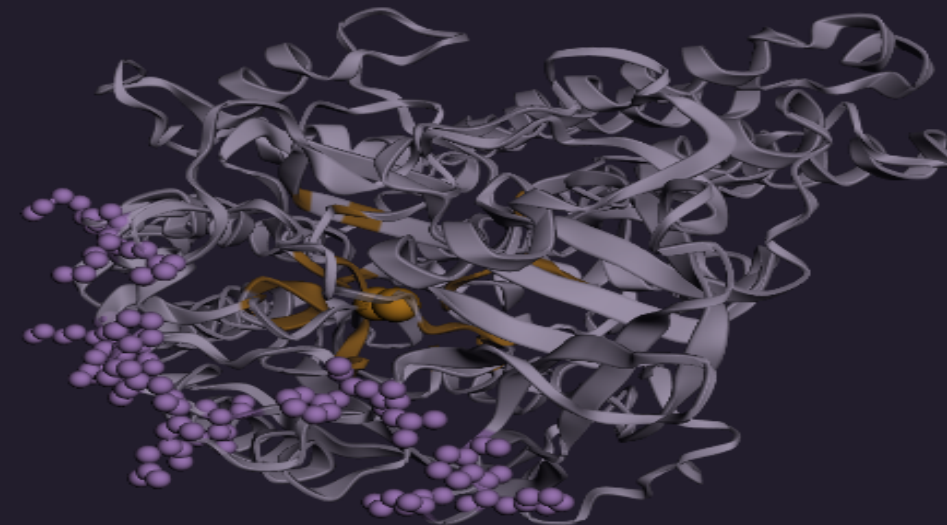
PSMA is the target that survives our triage of twenty antigens with a clinical programme: it is present in the patients who will metastasize and it spares healthy tissue. On its structure we map the surface residue by residue and propose where an antibody should bind — ruling out the active site and the face that sits against the membrane.

Phase one: epitope selection.

Where the molecule must bind, before building it.



Surface map



Candidate patch

Thank you

Questions

Data: GSE46691 (Mayo Clinic) · GSE116918 (Belfast) · TCGA-PRAD

Public retrospective cohorts · no prospective validation, no approval for clinical use